Accumulation of Potential Driver Genes with Genomic Alterations Predicts Survival in High-Risk Neuroblastoma

Chen Suo[1,†], Wenjiang Deng[2,†], Trung Nghia Vu[2], Leming Shi[1], Yudi Pawitan[2,*]

[1]Collaborative Innovation Center for Genetics and Development, Ministry of Education Key Laboratory of Contemporary Anthropology and the State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai, China; [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

[†]These authors contributed equally to this work; *Correspondence: yudi.pawitan@ki.se

Abstract

Background: Neuroblastoma is the most common pediatric malignancy with heterogeneous clinical behaviors, ranging from spontaneous regression to aggressive progression. Many studies have identified potential aberrations related to the pathogenesis and prognosis, but predicting tumor progression in and clinical management of high-risk patients remains a big challenge.

Method: We integrate gene-level expression, array-based comparative genomic hybridization and functional gene-interaction-network profile of 145 neuroblastoma patients to detect potential driver genes. The drivers are summarized within each patient into a score (DGscore), and we then validate its clinical relevance in terms of association with patient survival.

Results: Focusing on the subset of 48 clinically defined high-risk patients, we identify 193 recurrent copy number aberrations (CNAs), resulting in 274 altered genes with copy number gain or loss which have corresponding impact on the gene expression. Using a network enrichment analysis, we detect four common driver genes, *ERCC6*, *HECTD2*, *KIAA1279*, *EMX2*, and 66 patient-specific driver genes. Patients with high DGscore, i.e. carrying more copy-number-altered genes with correspondingly up or down-regulated expression and functional implications, have worse survival than those with low DGscore (P = 0.006). Furthermore, Cox proportional-hazards regression analysis indicates that, adjusted for age, tumor stage or *MYCN* amplification, DGscore is the only significant prognostic factor for high-risk neuroblastoma patients (P = 0.008).

Conclusions: Integration of genomic copy-number alteration, expression and functional interaction-network data reveals clinically relevant and prognostic putative driver genes in neuroblastoma. The identified putative drivers may give us new drug targets for individualized therapy.

Introduction

Neuroblastoma, an embryonal malignancy in sympathetic nervous system, is the most frequent extracranial solid tumor in children. It accounts for 7% of pediatric oncology and 15% of childhood cancer deaths. Neuroblastoma is highly heterogeneous with various clinical courses, ranging from spontaneous regression to aggressive and therapy-resistant progression despite intensive treatment [1-3]. Prognosis of neuroblastoma patients is associated with many factors, such as age at diagnosis, the International Neuroblastoma Risk Group (INRG) staging and oncogene *MYCN* amplification. Patients with stage 4 disease >18 months at diagnosis or patients of any age and stage with *MYCN*-amplified tumors are referred as high-risk patients [4]. Although several alterations including *MYCN* amplification, *TERT* rearrangements, *ALK* and A*TRX* mutations are identified to be associated with neuroblastoma, detection of potential drivers is still hampered by the low

mutation frequency and few recurrently mutated genes [5]. We hypothesize that additional structural alterations rather than point mutations might occur in high-risk neuroblastoma.

In this study, we aim to identify potential drivers of neuroblastoma by integrating various molecular features, including RNA sequencing (RNA-Seq), array comparative genomic hybridization (aCGH) data for copy number alterations (CNAs) and functional gene-interaction network. The drivers are defined as recurrent genomic alterations in tumor patients with consistent gene expression and with an important role in the functional interaction network. Furthermore, to assess the clinical relevance of the detected potential driver genes, we validate them in terms of association with patient survival. We demonstrate that the integration of diverse omics and functional data can provide more biologically and clinically relevant insight in neuroblastoma research in terms of potential drug targets.

Methods

Patients and datasets

The Neuroblastoma Data Integration Challenge of CAMDA 2017 provides expression profiles of 498 neuroblastoma patients, of which 145 patients have both RNA-Seq and aCGH data. There are 89 male and 56 female patients, and the age at initial pathological diagnosis ranged from 0 to 24.6 years old, with a median of 1.2 years old. Forty-eight out of 145 patients are clinically defined as high-risk neuroblastoma and 97 as low-risk [4]. The *MYCN* gene is a common proto-oncogene in neuroblastoma and examined by clinical diagnostic FISH test. We categorize the patients into 23 with *MYCN* amplification and 122 without *MYCN* amplification, respectively. Staging by the International Neuroblastoma Staging System (INSS), there are 33 patients at stage I, 20 at stage II, 20 at stage III, 47 at stage IV and 25 at stage IV-S. In order to optimize power, we shall focus our analysis to the 48 high-risk (HR) patients.

Integrative statistical analysis

Figure 1a presents an overview of the procedures to identify potential driver genes, including data pre-processing, copy number calling, integrative analysis and clinical validation.

First, we use two computational algorithms, MPSS [6] and cnvpack [7], to identify CNAs within and commonly across patients, respectively. MPSS takes a robust smooth segmentation approach to identify whether a segment is a true CNA [6]. We then use cnvpack to detect recurrent CNA regions, which is defined as alterations occurred in at least 10% of all patients [7]. To investigate the impact of CNAs on gene expression, we annotate genes on CNAs and compare the gene expression pattern in samples with alterations compared to samples with normal copy number. We keep genes which exhibit significantly over-expression with amplifications compared to the non-altered samples, based on p-value (P) < 0.05 using one-sided Welch's *t*-test, vice versa for genes with deletions. These genes are then chosen as potential drivers and referred as functional gene set (FGS, Fig. 1a). In parallel to the CNA analysis, we obtain gene expression data for 60,776 genes derived from RNA-Seq, which are measured in FPKM with corrections using Magic-AceView (MAV) pipeline [8]. The raw gene expression data are then centered and variance scaled within each patient. Since no paired normal tissues are available for the patients, we rank the expression level for extremity of each gene across the original 498 samples. For each patient, we then keep the top 200 highest ranked genes as patient-specific extremely expressed genes or the so-called patient-specific expression-altered gene sets as shown in our analysis pipeline (Altered Gene Set, AGS, Fig. 1a). The collection of recurrent patient-specific AGS is considered as common AGS. In addition to the expression profile-based AGS, 52 neuroblastoma-related genes from literature [9] are also considered as AGS.

Next, to integrate the results of copy number alteration and gene expression data, we implement network enrichment analysis (NEA) as follows. The key idea for NEA is that the functional impact of each copy-number-altered gene can be assessed according to the number of differentially expressed neighbors in a gene interaction network. We use a comprehensive network containing 1.4 million functional interactions between 16,288 HUPO genes/proteins [10]. Each copy-number-altered gene in FGS is assessed for its central

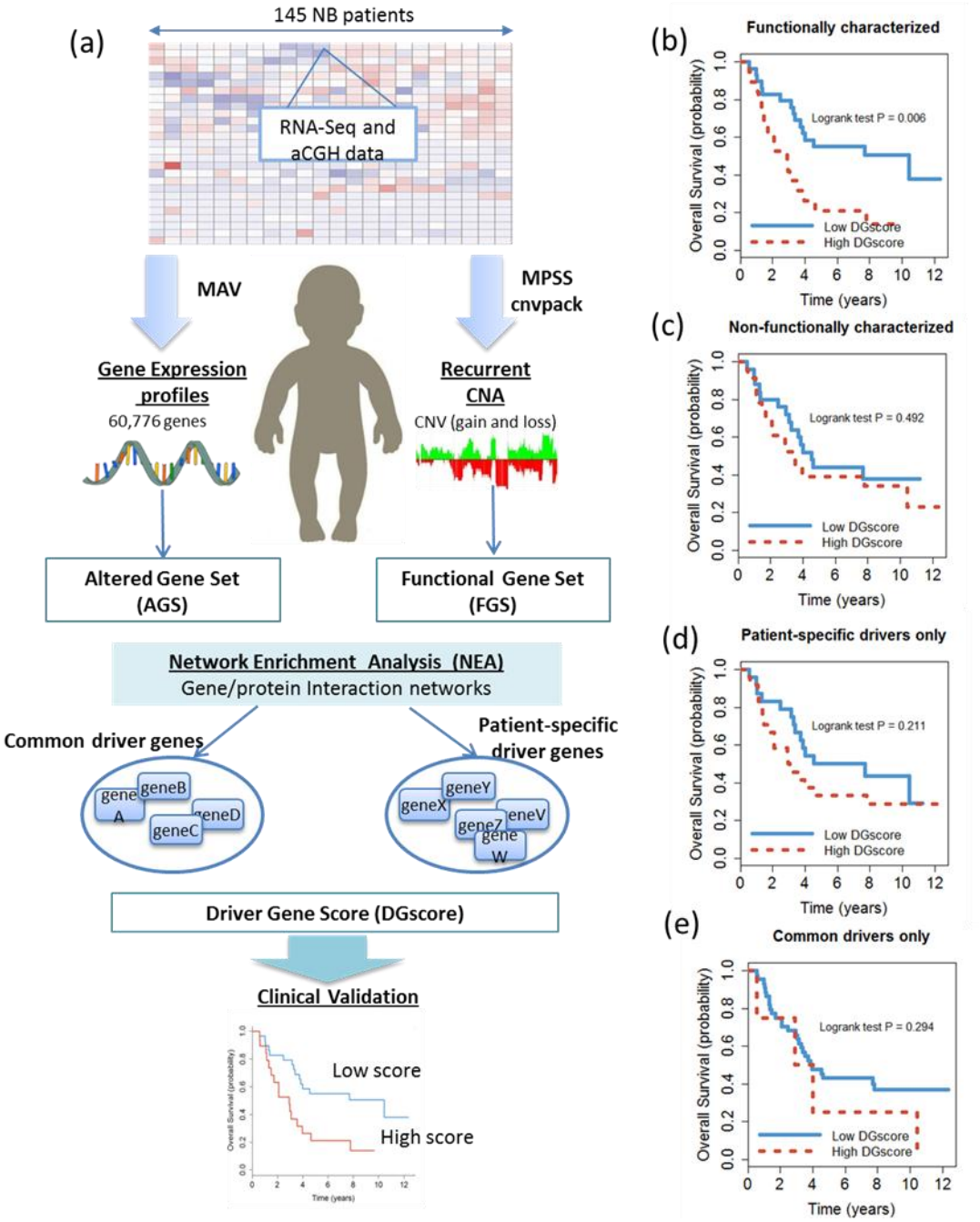

Figure 1. (a) Flowchart of the identification of potential driver genes and clinical validation. (b~e) Comparison of survival analysis for 48 high-risk patients split by different levels of omics integration.

functional role in modulating the expression of its interacting neighbors in the network. The significance is accessed using a quantitative enrichment score (z-score), which measures the over-representations of direct links between the AGS and FGS. Genes which are functionally significant, with z-score>2, are kept as

putative driver genes. We compute the total number of drivers with CNAs in each patient and term the number 'driver gene score' (DGscore). Finally, we use the DGscore to compare the prognosis of patients with DGscore larger than the median versus those lower than the median.

Results

Driver genes in high-risk neuroblastoma

Among 48 high-risk (HR) neuroblastoma patients, we identify 4,058 CNAs with an average 84 and range 9~433. Next, we detect 193 recurrent CNAs occurred in at least 5 (10%) of the 48 subjects. The recurrent CNAs contain a total of 6,390 genes. After filtering we have a final set of 274 recurrently altered genes which are served as FGS in the network enrichment analysis. To identify patient-specific driver genes, we perform the NEA analysis within each sample, where the AGS is the top 200 patient-specific extremely expressed genes and FGS is the patient-specific genes among the 274 altered genes. We detect 66 unique patient-specific drivers, with a median number of 2.8 in each high-risk neuroblastoma patient. To identify common driver genes, FGS and AGS are built as the follow. We apply a more stringent criterion by excluding recurrent CNA regions containing both amplifications and deletions among patients to create FGS. The reduced FGS contains 30 genes in which 10 genes exhibiting only amplifications and 20 genes with only deletions. Next, an AGS is derived from 52 candidate neuroblastoma genes from literature [9] and 111 common extremely expressed genes occurred in >4 patients. Finally, the NEA analysis finds four common potential driver genes *ERCC6*, *HECTD2*, *KIAA1279* and *EMX2*.

To examine the clinical relevance of the potential drivers, we divide 48 HR samples into high and low DGscore groups, where the high DGscore is defined as larger than the median value of the DGscore. Fig. 1b shows that neuroblastoma HR patients with a high DGscore have poor survival compared with low DGscore patients (Figure 1b, $P = 0.006$). However, if we simply summarize the 274 non-functionally characterized CNA genes, we would not be able to predict well the patients' survival (Fig. 1c, $P = 0.492$). This indicates the importance to functionally characterize recurrent altered genes by NEA. Another advantage of DGscore is that by integrating information of common driver genes and patient-specific driver genes, it can capture both the recurrent and individualized signatures in tumors. Separately using either only patient-specific driver genes (Fig. 1d) or only common driver genes (Fig. 1e) for NEA cannot predict patient survival well ($P > 0.2$).

Table 1. Cox proportional-hazard regression models of survival.

Model	Variable	Hazard ratio	P *
Model 1a	DGscore	2.69	0.008
Model 1b	Tumor stage	1.41	0.52
Model 1c	*MYCN* amplification	1.18	0.65
Model 1d	Age	1.00	0.058
Model 2	DGscore+tumor stage[a]		
	DGscore	2.69	0.008
	Tumor stage	1.41	0.52
Model 3	DGscore+*MYCN*[b]		
	DGscore	2.68	0.007
	MYCN amplification	1.15	0.70
Model 4	DGscore+age		
	DGscore	2.67	0.008
	Age	1.00	0.064

[a]Stage 4/4S are compared against Stage I-III

[b]No *MYCN* amplification is used as reference group

*P-values from the Wald test.

For neuroblastoma, tumor stage, *MYCN* oncogene amplification and age are known prognostics factors, but not necessarily so for HR patients. We thus investigate whether the DGscore has a prognostic value independent of the previously known predictors. To do that, we include these factors in Cox regression analysis of HR patients. In Table 1, the first four models display the individual predictors in univariate regression, where DGscore is the only significant predictor (Model 1a, P=0.008). Note, in particular, that *MYCN* amplification is not significant (Model 1c, P=0.65). The last three models show that DGscore remains highly significant after adjusting for tumor stage, *MYCN* amplification or age (Models 2, 3 and 4).

Discussion and Conclusion

We have implemented an integrative omics analysis to identify potential driver genes in neuroblastoma and validate these drivers clinically in terms of survival prediction. The results show that high-risk neuroblastoma patients who carry more copy-number-altered genes with functional implications and extreme expression patterns have worse survival than those with less potential driver genes. The potential drivers, especially the patient-specific drivers, may provide potential drug targets for individualized precision medicine and new information in understanding the tumor biology.

We also perform NEA analysis for the whole 145 neuroblastoma patients. No common driver genes are detected for the whole 145 samples, perhaps due to the diverse clinical outcome of neuroblastoma. Interestingly, our patient-specific analysis successfully identifies individualized drivers and clearly separates the patients into two distinct survival groups (results not shown). Some identified common driver genes in HR patients has been discovered to play important roles in neuronal differentiation in previous studies. For example, *ERCC6*-depleted neuroblastoma cells show defects in gene expression programs required for neuronal differential and fail to differentiate and extend neurites [11]. *EMX2* is a prognostic and predictive biomarker in malignant pleural mesothelioma [12]. Nonsense mutations in *KIAA1279* are also associated with malformation of the central and enteric nervous system [13]. Furthermore, the top two mostly recurrent drivers identified through the patient-specific approach, *OTOP3* and *MYCN*, are considered as a driver event in 13 (27%) out of the 48 HR patients. Herein, *MYCN* is one of the best characterized genetic alterations in neuroblastoma; and copy number gain of chromosome 17q, where *OTOP3* locates, is a known neuroblastoma risk factor.

In our future work we would use the transcript-level expression data from RNA-Seq to refine the current driver signatures. In addition, patients whose expression profiles are available but not the aCGH data may be used to partially validate the discovered drives. One limitation of our current analysis is the small data size. We would need an independent dataset with both aCGH and expression data for further validation.

Reference

1.Molenaar JJ *et al. Nature* 2012, 483(7391):589-593.

2.Maris JM *et al. Lancet* 2007, 369(9579):2106-2120.

3.Pugh TJ *et al. Nature Genetics* 2013, 45(3):279-284.

4.Zhang W *et al. Genome Biology* 2015, 16:133.

5.Peifer M *et al. Nature* 2015, 526(7575):700-704.

6.Teo SM *et al. Bioinformatics* 2011, 27(11):1555-1561.

7.Mei TS *et al. BMC Bioinformatics* 2010, 11:147.

8.Thierry-Mieg D *et al. Genome Biology* 2006, 7 Suppl 1:S12 11-14.

9.Cao Y *et al. Oncotarget* 2017, 8(11):18444-18455.

10.Alexeyenko A *et al. BMC Bioinformatics* 2012, 13:226.

11.Wang Y *et al. PNAS* 2014, 111(40):14454-14459.

12.Giroux Leprieur E *et al. Lung Cancer* 2014, 85(3):465-471.

13.Brooks AS *et al. American Journal of Human Genetics* 2005, 77(1):120-126.